



códigofacilito

---

# CódigoFacilito.

Introducción a procesamiento de lenguaje natural (PLN)

Dr. Niels Martínez Guevara



## 1 Características del lenguaje

Lo que no es lenguaje humano

Ley de Zipf

Introducción a la ambigüedad

## 2 ¿Qué es PLN?

## 3 Transformación de los datos

Palabras funcionales

Morfemas



# Introducción



Figura: Diagrama de Dikw enfocado en ciencia de datos <sup>1</sup>

---

<sup>1</sup> J. D. Kelleher y B. Tierney (2018). *Data science*. MIT Press



# ¿Ciencia de Datos? ¿Machine Learning? o ¿Inteligencia Artificial?

## Fenómeno

Entendemos como fenómeno todo aquello que podemos percibir por nuestros sentidos y la razón<sup>2</sup>. Cuyas características podemos registrar en variables.

## Datos estructurados

Son aquellos que se asocian con una variable por ejemplo velocidad, color, sabor, etc. Estas pueden ser cualitativas (descriptivas u ordinales) y cuantitativas<sup>3</sup>.

## Datos no estructurados

Son aquellos que no poseen una estructura generalizable, como los **textos**, las imágenes, videos, audios, etc<sup>4</sup>.

<sup>2</sup> I. Kant (1978). «Crítica a la razón pura. Prólogo a la segunda edición». En: Madrid. Alfaguara

<sup>3</sup> J. D. Kelleher y B. Tierney (2018). *Data science*. MIT Press

<sup>4</sup> J. D. Kelleher y B. Tierney (2018). *Data science*. MIT Press



# Introducción al lenguaje como fenómeno

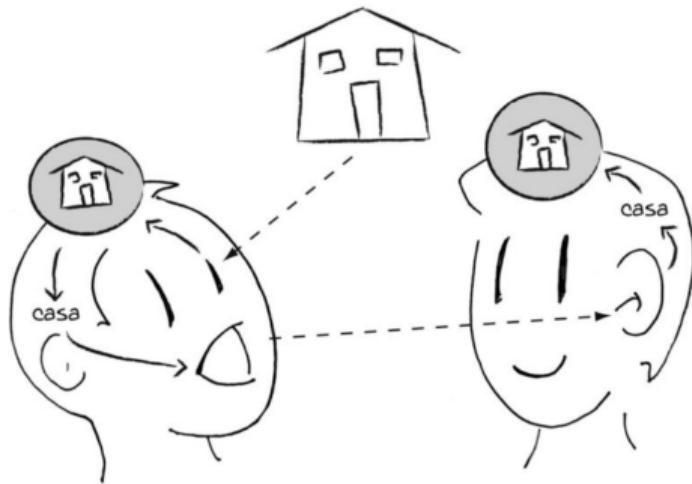


Figura: Representación del proceso de comunicación<sup>5</sup>

<sup>5</sup> J. Muñoz-Basols, N. Moreno, T. Inma y M. Lacorte (2016). *Introducción a la lingüística hispánica actual: teoría y práctica*. Routledge



# ¿Qué es la lingüística?, ¿El lenguaje? y ¿por qué todo esto supone un problema?

- Formalmente, un lenguaje se puede considerar como un conjunto de cadenas sobre un alfabeto.
- Podemos ser comprendidos por otros que conocen el mismo lenguaje.
- Un lenguaje humano es un sistema de comunicación estructurado usado por un grupo social.
- Producimos cadenas de símbolos que denotan información.
- Los lingüistas estudian el lenguaje como algo que se manifiesta, en ningún momento se dedican a normalizarlo, simplemente registran su comportamiento. Si la RAE no te puede decir como hablar, solo registra como lo hace un grupo social. Por lo que hablar en chileno esta bien.



# El lenguaje es complejo

Steven Pinker

La capacidad de comunicarnos nos resulta tan natural y cotidiana, que tendemos a pasar por alto lo compleja y asombrosa que es<sup>6</sup>.

Noah Chomsky

El lenguaje es algo tan apgado a la experiencia humana que lo podemos comparar como un sistema biológico, como la capacidad de respirar<sup>7</sup>.

---

<sup>6</sup> S. Pinker (2003). *The language instinct: How the mind creates language*. Penguin uK

<sup>7</sup> N. Chomsky (1998). «Nuestro conocimiento del lenguaje humano: perspectivas actuales». En: Santiago de Chile: Ed. Universidad de Concepción & Bravo y Allende



# ¿Qué es lo correcto? ¿se puede normar? La lengua como fenómeno social

Cuando estudiamos el lenguaje es de suma importancia, conocer las características de la lengua que estamos estudiando. Es por ello que utilizaremos un enfoque basado en corpus.

## Corpus

Un corpus es una colección de ejemplos de lenguaje cómo ocurren naturalmente. Los ejemplos deben ser **colectados correctamente**, y deben encontrarse almacenados en un soporte digital.

¿Colectados correctamente? Estos dependerán del objetivo de la investigación. De nada sirve si tengo un corpus de España si la aplicación que deseo implementar es en México (o un contexto específico).



# Gramática y variantes en el lenguaje

- Gramática es un término que se usa de forma ambigua.
- En el sentido más simple es: el conocimiento lingüístico que tenemos sobre las unidades y reglas del lenguaje:
  - para producir sonidos (fonología)
  - para formar palabras (morfología)
  - para combinar palabras en oraciones (sintaxis)
  - para asignar significado (semántica)
- La gramática y nuestro diccionario mental de palabras (léxico) representan nuestra competencia lingüística.



## Gramática y variantes en el lenguaje (II)

- Todo humano que habla un lenguaje sabe su gramática.
- Cuando los lingüistas describen un lenguaje intentan hacer explícitas las reglas en la cabeza de los hablantes.
- Hay diferencias, pero hay reglas compartidas.
- Las reglas compartidas permiten la comunicación.



## Gramática y variantes en el lenguaje (III)

Esta descripción es un modelo de la competencia lingüística de los hablantes. Se conoce como gramática descriptiva.

### Ojo

No dicta como hablar, describe cómo se usa el lenguaje.

- Si la gramática descrita difiere de la propia competencia lingüística no implica que hablemos mal.
- No hay lenguajes superiores o inferiores.
- Toda gramática interna es igualmente compleja e igualmente expresiva.
- Todo lo que se puede expresar en un lenguaje, se puede expresar en otro (incluidas las LS).



## Gramática y variantes en el lenguaje (IV): A New Hope

- Existen y han existido “puristas” del lenguaje: personas que aseveran que hay formas “correctas” e “incorrectas” del lenguaje. Pretenden prescribir mediante la gramática, no describir.
- Estos son fenómenos sociales.
- La evidencia histórica sugiere que la concepción del lenguaje “correcto” suele ser el de las élites y el “incorrecto” el de los grupos oprimidos.
- Ej. Revolución francesa, el latín vulgar y ñeros vs fresas.



## Lo que no es lenguaje humano

La mayoría de los seres vivos poseen sistemas de comunicación.

- Comunicación sonora y gestual.
- Señales de estrés, apareamiento, etc.
- La evidencia muestra que estos sistemas carecen de creatividad.
- No son sistemas discretos.



## Lo que no es lenguaje humano

- Los humanos usamos símbolos discretos: gramática y léxicos finitos. Estos se combinan para crear cadenas infinitas.
- En la combinación radica esa creatividad.
- Los animales usan llamadas finitas. Hay un catálogo finito de llamadas o comportamientos. Cada uno señala una necesidad específica: hambre, miedo, etc.
- En laboratorio se logran reproducir las señales de varios animales.

Bertrand Russell

Un perro puede ladrar con elocuencia, pero no puede decir que sus padres eran honestos aunque pobres.



## Lo que no es lenguaje humano

- Durante el siglo XX se hicieron varios experimentos. Sobre todo con primates. En ninguna instancia un primate pudo articular lenguaje al nivel de un niño humano.
- Podían asociar vocabulario fijo a cosas concretas. e.g. aprender que “galleta” es comida.
- Ninguno pudo juntar símbolos para formar frases.
- Usando signos de ASL con primates, a lo más se observaron instancias de uso de dos símbolos.
- Ningún animal ha aprendido nunca LS.
- Desde una perspectiva estadística (o distribucional) el lenguaje humano posee comportamientos diferentes al de los animales<sup>8</sup> o a la generación de cadenas aleatorias<sup>9</sup>.

---

<sup>8</sup> R. Suzuki, J. R. Buck y P. L. Tyack (2005). «The use of Zipf's law in animal communication analysis». En: *Animal Behaviour* 69.1, F9-F17

<sup>9</sup> R. Ferrer-i-Cancho y B. Elvevåg (2010). «Random texts do not exhibit the real Zipf's law-like rank distribution». En: *PLoS One* 5.3, e9411



# Características estadísticas: Ley de Zipf

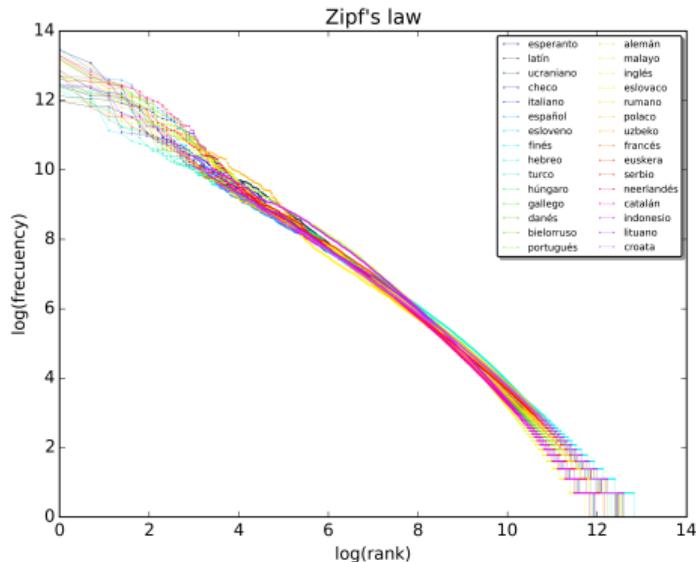


Figura: Gráficas de ranking versus la frecuencia de las primeras 10 millones de palabras en 30 Wikipedias (descargas de octubre del 2015) <sup>10</sup>

<sup>10</sup> Wikipedia (2023). Ley de Zipf — Wikipedia, The Free Encyclopedia. <http://es.wikipedia.org/w/index.php?title=Ley%20de%20Zipf&oldid=148265153>. [Online; accessed 13-November-2023]



## Ley de Zipf

La llamada ley de Zipf, formulada en la década de 1940 por George Kingsley Zipf, lingüista de la Universidad de Harvard, es una ley empírica según la cual en una determinada lengua la frecuencia de aparición de distintas palabras sigue una distribución que puede aproximarse por:

$$P_n \sim 1/n^a \quad (1)$$

Donde  $P_n$  representa la frecuencia de la  $n$ -ésima palabra más frecuente y el exponente  $a$  es un número real positivo, en general ligeramente superior a 1. Esto significa que el segundo elemento se repetirá aproximadamente con una frecuencia de  $1/2$  de la del primero, el tercer elemento con una frecuencia de  $1/3$  del primero y así sucesivamente<sup>11</sup>.

---

<sup>11</sup> M. A. Montemurro (2001). «Beyond the Zipf–Mandelbrot law in quantitative linguistics». En: *Physica A: Statistical Mechanics and its Applications* 300.3-4, págs. 567-578



# ¿Para qué tantos conceptos lingüísticos?

Como computólogos (o estadísticos) interesados en procesar lenguaje:

- Nos permite intuir dónde están los límites de lo que podemos hacer.
  - ¿Qué estamos buscando?
  - ¿Cómo podemos encontrarlo?
- Desde una perspectiva computacional tenemos ventajas y desventajas.



## Ventajas

- Cadenas infinitas pero símbolos finitos.
- Con suficientes ejemplos podemos colectar oraciones/palabras/signos que tienen alta probabilidad de ocurrir.
- Las reglas son complejas pero finitas.
- Con suficiente poder de cómputo/observación podemos inferir reglas a partir de los datos.
- Existen límites (biológicos/sociales) que reducen el tamaño del conjunto posible de cadenas.
- En el peor caso podemos predecir qué cadenas se van a utilizar (fuerza bruta).
- Asumiendo que tenemos suficiente información y hacemos las cosas bien.



## Deventajas

- La creatividad de los hablantes: Incertidumbre.
- Las reglas y símbolos no son fijos. Hay variación.
- Hay reglas que rigen los cambios pero son más difíciles de procesar.
- Se requieren muchos ejemplos.
  - Ethnologue registra ≈ 7100 lenguajes. Lo que se clasifica como un sólo lenguaje tiene muchos dialectos.
  - Más de 6000 son hablados por menos de 1 millón de personas.
  - 40% son hablados por menos de 1000 personas (en extinción).
  - Una gran parte están poco documentados.
  - Una gran parte no tienen sistema de escritura



# Tipos de ambigüedad

- Ambigüedad fonética y fonológica
- Ambigüedad léxica
- Ambigüedad sintáctica
- Ambigüedad pragmática
- Anáfora



## Ambigüedad fonética y fonológica

Es cuando la articulación de un enunciado se puede interpretar de más de una forma (homonimia).

- tuvo / tubo
- a ver / haber
- allá / Haya
- Yo lo coloco y ella lo quita.



# Ambigüedad fonética y fonológica

Los cambios articulatorios pueden ser muy pequeños.

- b/p: peso → beso
- k/g: cango → ganso
- t/d: teme → deme
- r/l: pera → pela
- Los instrumentos de captura de muy baja calidad.
- Modelos o datos insuficientes.
- e.g. variación de los hablantes (pronunciación distinta).



# Ambigüedad fonética y fonológica

—Qué coincidencias tiene la vida. Tú te llamas Solovino y yo tomo sólo vino.

—Suéltame, señor.



Figura: Ejemplo de ambigüedad fonética



## ¿Cómo se ha resuelto?

- Entrenamiento estadístico
  - Se colecta un corpus.
  - Se etiqueta con información lingüística.
  - Se analiza: distribución de fonemas y enunciados más comunes.
  - Permite al sistema adivinar (educated guess).
- En sistemas específicos (Alexa, Google Home, Cortana) se tiene un modelo híbrido.
  - Hay un conjunto de frases que pueden ocurrir.
  - El sistema intenta empatar los sonidos con las frases.
- Entrenamiento personalizado
- Requiere muchos recursos y conocimiento de la población objetivo.



## Ambigüedad léxica

Es cuando el significado de una palabra se puede interpretar de más de una forma.

- banco (institución/peces/silla)
- carta (correspondencia/naipe/menú)
- planta (industria/ser vivo)
- jaguar (auto/animal)



## Ambigüedad léxica

**EL PAÍS**  
**ARCHIVO**

EDICIÓN  
IMPRESA

Hemeroteca ▾

DOMINGO, 11 de mayo de 2003

---

REPORTAJE: OFERTAS DE EMPLEO

### *Casi la mitad de los jóvenes son temporales*

La tasa se ha reducido desde que gobierna el PP, pero aumenta en los últimos años en Madrid y Barcelona

Figura: Ejemplo de ambigüedad léxica



## Ambigüedad léxica



Figura: Ejemplo de ambigüedad léxica



# Ambigüedad léxica

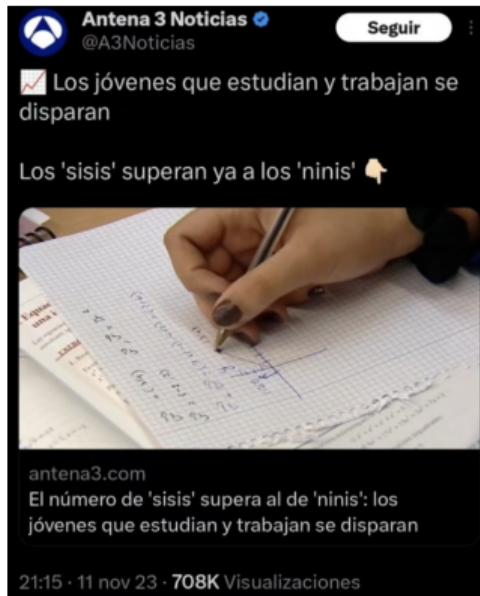


Figura: Ejemplo de ambigüedad léxica



# Colocaciones

Palabras que, juntas, adquieren un significado distinto.

- Tercera edad
- Fuego amigo
- Para adultos
- Reajuste de personal

Para un sistema de PLN esto no es evidente.



## Ambigüedad léxica

el profe: tranquilos, el examen será fácil

el examen: señale cual de los dos es un toro mecánico



Figura: Momazo



## ¿Cómo se ha resuelto?

- Métodos basados en conocimiento.
  - Diccionario
  - Redes semánticas
- Entrenamiento estadístico
  - No supervisado (e.g. agrupación de co-ocurrencias).
  - Semi-supervisado (e.g. semántica distribucional).
  - Supervisado (corpus anotado con etiquetas semánticas).
- Usualmente se usa una mezcla de ambos enfoques.
- Requiere muchos recursos y conocimiento de la población objetivo.



## Ambigüedad sintáctica

Cuando la estructura de un enunciado se puede interpretar de maneras distintas. No se sabe a que refiere cada parte del enunciado.

### Ejemplo

Vi al hombre en la colina con un telescopio.



## Ambigüedad sintáctica



Figura: Vi al hombre en la colina con un telescopio.

## Ambigüedad sintáctica

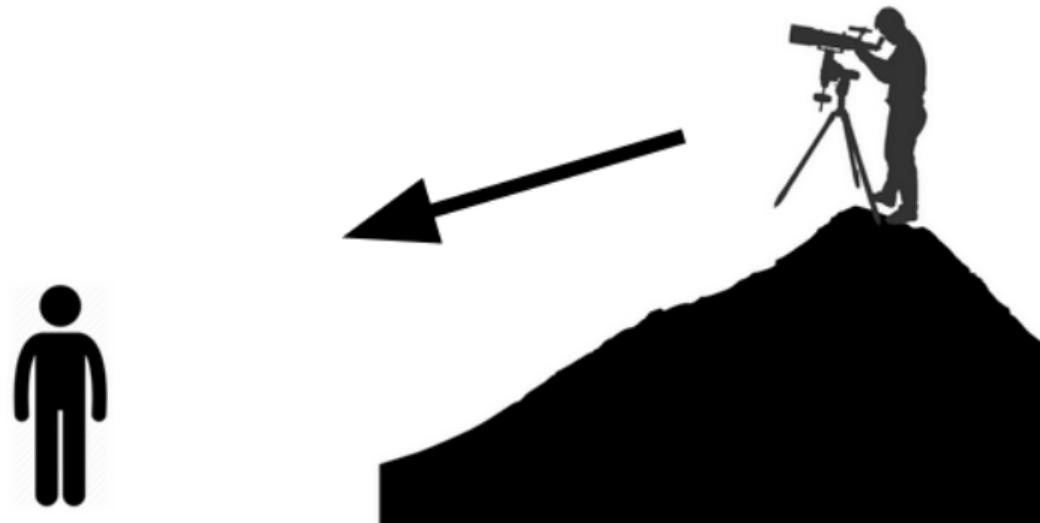


Figura: Vi al hombre en la colina con un telescopio.

## Ambigüedad sintáctica

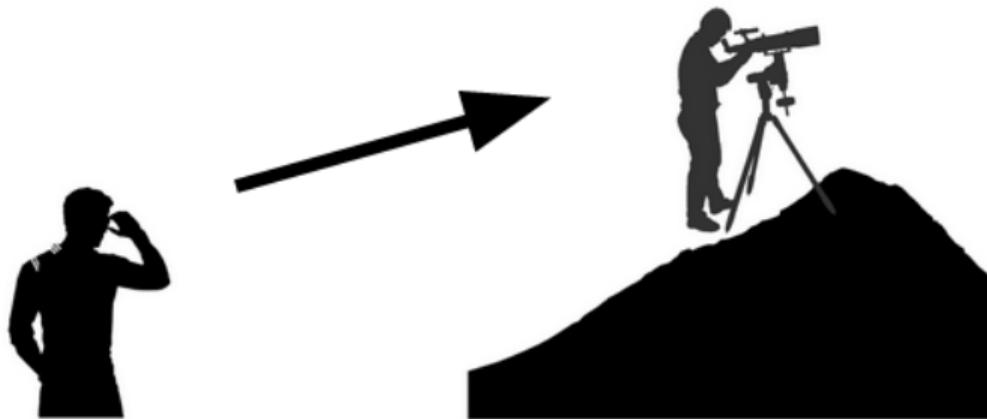


Figura: **Vi** al hombre en la colina con un telescopio.

## Ambigüedad sintáctica



Figura: **Vi al hombre en la colina con un telescopio.**

## Ambigüedad sintáctica



Figura: **Vi al hombre en la colina con un telescopio.**

# Un Boeing 747 regresa a Madrid por una grieta

MADRID.— Un Boeing 747 de [REDACTED], que realizaba ayer el trayecto entre Madrid y Nueva York, regresó al aeropuerto de Barajas aproximadamente una hora y media después de despegar, tras detectarse una leve grieta en el cristal de la ventanilla del copiloto, que no llegó a romperse, según informa Efe.

El vuelo 6251 despegó desde Barajas prácticamente a la hora prevista (13.10) y unos 50 minutos después se descubrió una leve hendidura en uno de los cristales de la cabina de los pilotos.

Aunque el avión podía continuar su vuelo hacia Nueva York sin complicaciones, el comandante decidió

Figura: Ambigüedad en la frase preposicional



## ¿Cómo se ha resuelto?

- Entrenamiento Sobre corpus de árboles de derivación. e.g. Algoritmo de Nivre
- Word embeddings.
- Aproximaciones híbridas con información léxica.
- Descripción manual de reglas sintácticas.



# Ambigüedad pragmática

Es cuando la interpretación de un enunciado depende del contexto de la enunciación.

¿Cómo te doy lo que no tengo?

- Con un ladrón (no traigo dinero).
- En el trabajo (no he terminado).
- En una película de Pedro Infante (soy pobre).
- En Game of Thrones o algún harem (soy eunuco).



## Anáfora

Uso de una palabra para hacer referencia o sustituir otra usada previamente en el discurso, usualmente para evitar su repetición.

- Pronominal: Juan encontró al amor de **su** vida.
- Definitud de frase nominal: **La relación** no duro mucho.
- Quantificación/Ordinal: No fue la **última**, pronto comenzó una nueva.



## ¿Cómo se ha resuelto?

- Restricciones de eliminación
- Se compara la anáfora con los posibles referentes.
- Se descartan discordancias. e.g. "una" sustituye un referente singular - femenino;
- Asignación por pesos.
  - Se asignan valores entre cada anáfora y los posibles referentes.
  - Criterios de proximidad, centralidad, rol semántico/sintáctico.
- Se emparejan las que tienen valores más fuertes: se asume que son las más probables.



# Conclusión

- La ambigüedad es un problema central del PLN.
- La gran mayoría de los sistemas tienen que tratarla.
- Se requiere conocimiento lingüístico para resolverla.
- Está presente a todos los niveles de estudio.
- Normalmente al mismo tiempo.
  - Una ambigüedad se puede arrastrar a través de todos los niveles:
  - "Yo lo coloco y ella lo quita."



## 1 Características del lenguaje

Lo que no es lenguaje humano

Ley de Zipf

Introducción a la ambigüedad

## 2 ¿Qué es PLN?

## 3 Transformación de los datos

Palabras funcionales

Morfemas



# ¿Lingüística computacional o procesamiento de lenguaje natural?

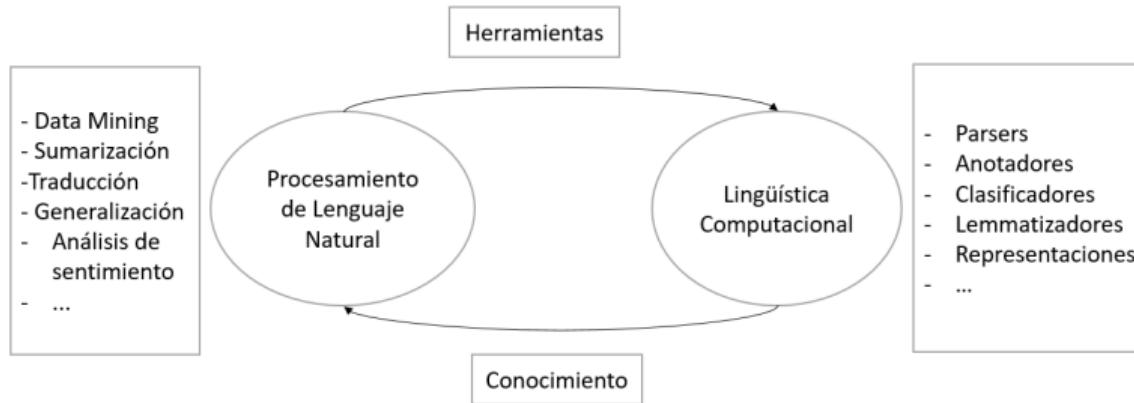


Figura: Relación entre la lingüística computacional y el procesamiento de lenguaje natural<sup>12</sup>

<sup>12</sup> D. Jurafsky y J. H. Martin (s.f.). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.



# Inicios de interacción



**Figura:** Juguete de 1920 que interactuaba con la palabra REX (reconociendo la frecuencia de 500Hz) correspondiente a la palabra <sup>13</sup>.

---

<sup>13</sup> D. Jurafsky y J. H. Martin (s.f.). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.

- Entrada de lenguaje ≈ “datos no estructurados”.
  - Sí tienen una estructura: las reglas/unidades del lenguaje.
- Las tareas de PLN intentan encontrar dicha estructura, ya sea:
  - Codificando reglas manualmente (algoritmos simbólicos).
  - Deduciendo reglas de los datos (algoritmos de aprendizaje).



- El PLN está relacionado con la IA: es una relación similar a la que tenemos con la Lingüística Computacional.
- La IA pretende construir máquinas con las mismas habilidades que los humanos. La habilidad de producir/entender lenguaje está incluida (Test de Turing)
- La mayoría de las tareas de PLN se podrían clasificar como “IA débil” .



- El problema principal de la IA es el contexto.
- El contexto es vasto.
- Los humanos adquieren y modifican su información contextual constantemente.
- Los sistemas de razonamiento son tan buenos como lo que saben.
- Es el mismo problema que tiene el PLN.



- El PLN también requiere contexto.
- La botella no cabe en la maleta porque es muy [pequeña/grande].
- Se requiere razonamiento espacial para desambiguar a quien se refiere “es”.
- Implica una relación fuerte entre la IA y el lenguaje.
- Sin embargo, PLN no es una prioridad para la IA.
- El PLN no es Natural Language Understanding (NLU).



## Tareas más comunes de PLN

- Recuperación de la información.
- Resolución de preguntas.
- Reconocimiento de entidades nombradas.
- Extracción de relaciones semánticas.
- Similitud textual.
- Etiquetado de partes de la oración.
- Clasificación de documentos.
- Extracción de terminología.
- Modelado de tópicos.



## 1 Características del lenguaje

Lo que no es lenguaje humano

Ley de Zipf

Introducción a la ambigüedad

## 2 ¿Qué es PLN?

## 3 Transformación de los datos

Palabras funcionales

Morfemas



# Algoritmos simbólicos y de aprendizaje

- Le mandamos datos a la computadora para que encuentre patrones.
- De preferencia etiquetados.
- Se crean modelos de forma automática.
- Regresión textual



## Creación de matrices términos - documentos

- Convertimos el texto a un modelo vectorial.
- Aplicamos una técnica de reducción de dimensionalidad (e.g. PCA).
- Entrenamos un modelo de regresión multiple con los componentes principales <sup>14</sup>.

---

<sup>14</sup> D. P. Foster, M. Liberman y R. A. Stine (2013). «Featurizing text: Converting text into predictors for regression analysis». En: *The Wharton School of the University of Pennsylvania, Philadelphia, PA*



## Matriz términos y documentos

- Representamos cada documento como un vector en un espacio determinado.
- Las componentes representan características que codificamos numéricamente.
- En el caso más sencillo representan la frecuencia de cada palabra (bolsa de palabras).
- La colección se agrupa en una matriz.



## Matriz términos y documentos

	$T_1$	$T_2$	$\dots$	$T_N$
$D_1$				
$D_2$				
:				
$D_n$				

Figura: Tabla de términos y documentos



## Matriz términos documentos

- D1 = Donde dice “dice” debe decir “debe decir”. Donde dice “debe decir” debe decir “dice”
- D2 = Donde dice “dice”
- D3= Aquí dice



## Matriz términos y documentos

	Donde	dice	debe	decir	Aquí
$D_1$	2	4	4	4	0
$D_2$	1	2	0	0	0
$D_3$	0	1	0	0	1

Figura: Tabla de términos y documentos



# Frecuencias

Las componentes pueden cuantificarse en N o R.

- Bolsa de palabras
- Frecuencia
- Tf\*Idf



# Términos

El vocabulario puede ser:

- Palabras individuales
- Grupos de palabras
- Palabras con cierta función (e.g. sustantivos)



# Palabras funcionales y de contenido

- Palabras de contenido

- Verbos
- Sustantivos
- Adjetivos
- Adverbios

- Palabras funcionales.

- Conjunciones
- Adposiciones
- Artículos
- Pronombres



## Palabras funcionales (Stop Words)

- No tienen significado léxico claro: sirven para especificar relaciones gramaticales.
  - “un niño” vs. “el niño”
  - “con hijos” vs. “sin hijos”
- Se suelen llamar palabras de clase cerrada.
- Es un conjunto estable.
- Es difícil que nuevas palabras funcionales entren al lenguaje.
  - observen los esfuerzos de introducir al español pronombres de género neutros.



## Palabras de contenido

- Denotan conceptos tales como objetos (alumno), acciones (estudiar), atributos (diariamente) e ideas (conceptos).
- Se suelen llamar palabras de clase abierta.
- Añadimos constantemente nuevas palabras a la clase.
  - Facebook (sustantivo)
  - googlear (verbo)



# Morfemas

- Es una unidad mínima de significado al interior de una palabra.
- Los morfemas se combinan para formar palabras.
- Esa combinación está regida por reglas.
  - constitucionalin, adaptadoin, competentein, completoin, cobrablein
  - No son palabras gramaticales en castellano.



# Morfología

- Estudia la estructura interna de las palabras (morfemas).
- Sus reglas de combinación como elementos discretos.
- Elementos que adquirimos subconscientemente.
- En si misma la palabra está compuesta por dos morfemas.
  - morf (relativo a la forma) y -ología (relativo al estudio)



## Raíces, Gramemas y bases léxicas

- En tareas como la regresión textual, modelado de tópicos y otras desde la perspectiva distribucional: es usual intentar reducir el número de gramemas.
  - stemming (reducir a una base léxica mínima)
  - lematizar (reducir a una palabra que represente a la raíz)
- Es usual remover las palabras funcionales.
- Es un esfuerzo de reducción de dimensionalidad.



# Stemming

- Los algoritmos de stemming intentan reducir la base léxica hasta una raíz.
  - niña → niñ
  - niño → niñ
  - niñez → niñ
  - niñas → niñ
  - niños → niñ
- Algoritmo de Porter (Snowball stemmer)
- Remoción de afijos conocidos



## Stemming

- Si tenemos un documento representado por el conjunto de términos  $w$ :
- $w = \{\text{la, niña, el, niño, y, la, niñez, de, las, niñas, y, los, niños}\}$
- $\text{no stopwords}(w) = \{\text{niña, niño, niñez, niñas, niños}\}$
- Colapsó trece dimensiones de la matriz en cinco (61 %).
- $\text{stem}(\text{no stopwords}(w)) = \{\text{niñ}\}$
- Colapsó cinco dimensiones de la matriz en una sola (92 %). Para un algoritmo de aprendizaje es una enorme reducción del costo computacional.



## Lematización

- Reduce los tipos a una forma “estándar” (lema). i.e. su “forma de diccionario”
- El lema representa la carga semántica de la raíz.
- Requiere más conocimiento lingüístico que el stemming. Niels está barqueando → barquear.
- Requiere encontrar la función gramatical.
  - robotizar → robot
  - robótico → robot
  - robotizado → robot
  - robóticamente → robot
- La forma estándar conserva el significado de la raíz.
- En este caso, el morfema libre “robot” .



## TF-IDF (Term Frequency - Inverse Document Frequency)

Es una medida numérica que expresa que tan relevante es un término (palabra) en un documento, dentro de un corpus.

$$tf(t, d) = \frac{\#t \in d}{\#PALABRAS \in d} \quad (2)$$

$$df(t) = \#t \in CORPUS \quad (3)$$

$$idf(t) = \frac{n}{df(t)} \quad (4)$$



## TF-IDF

Sin embargo, en corpus de gran tamaño IDF se vuelve inmanejable, para evitar esto se aplica el log de idf, para evitar el hecho de que una palabra no esté presente en el documento, se aplica la siguiente fórmula:

$$idf(t, d) = \log\left(\frac{n}{(idf + 1)}\right) \quad (5)$$

$$TF - IDF(t, d) = tf(t, d) * idf(t, d) \quad (6)$$



# Word Embeddings

## Skip-gram (relación entre palabras vecinas)

Funciona bien con una pequeña cantidad de datos de entrenamiento, representa bien incluso palabras o frases raras.

## CBOW (relación de ocurrencia en el corpus)

Varias veces más rápido de entrenar que el skip-gram, precisión ligeramente mejor para las palabras frecuentes<sup>15</sup>.

---

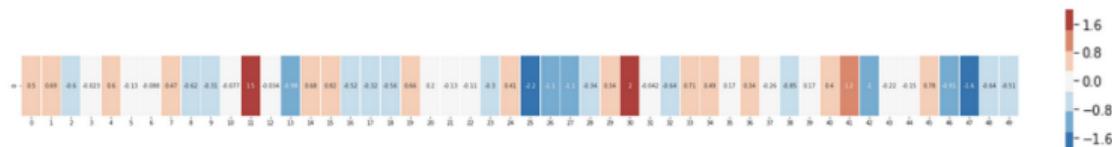
<sup>15</sup> [J. Brownlee \(ago. de 2019\). What are word embeddings for text?](#)



# Word 2 Vec

Supongamos que a partir de los artículos de wikipedia obtenemos el siguiente vector para la palabra "KING"<sup>16</sup>

```
[ 0.50451 , 0.68607 , -0.59517 , -0.022801, 0.60046 , -0.13498 , -0.08813 , 0.47377 , -0.61798 , -0.31012 ,
-0.076666, 1.493 , -0.034189, -0.98173 , 0.68229 , 0.81722 , -0.51874 , -0.31503 , -0.55809 , 0.66421 , 0.1961 ,
-0.13495 , -0.11476 , -0.30344 , 0.41177 , -2.223 , -1.0756 , -1.0783 , -0.34354 , 0.33505 , 1.8927 , -0.04234 ,
-0.64319 , 0.71125 , 0.49159 , 0.16754 , 0.34344 , -0.25663 , -0.8523 , 0.1661 , 0.40102 , 1.1685 , -1.0137 ,
-0.21585 , -0.15155 , 0.78321 , -0.91241 , -1.6106 , -0.64426 , -0.51042 ]
```



<sup>16</sup> J. Alammar (s.f.). *The illustrated word2vec*.



## Word 2 Vec

“king”



“Man”

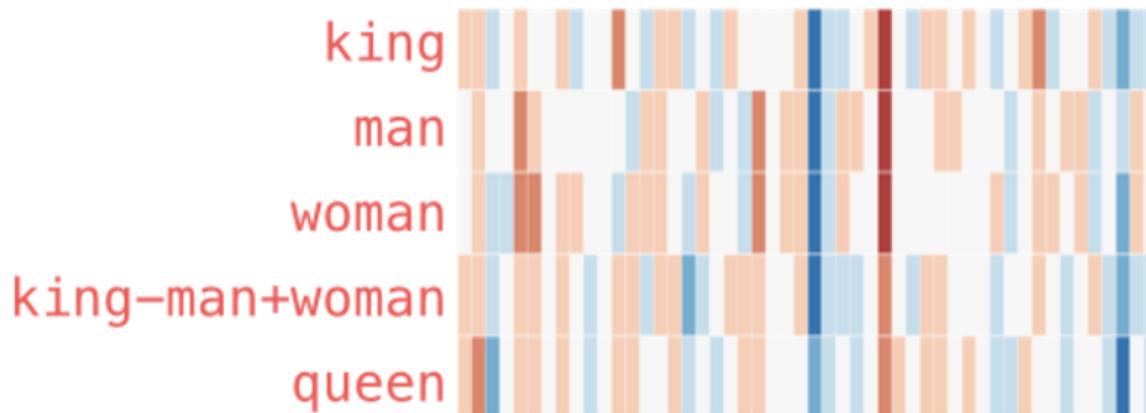


“Woman”



## Word 2 Vec

king - man + woman  $\approx$  queen



# Bibliografía I

Alammar, J. (s.f.). *The illustrated word2vec.*

Brownlee, J. (ago. de 2019). *What are word embeddings for text?*

Chomsky, N. (1998). «Nuestro conocimiento del lenguaje humano: perspectivas actuales». En: *Santiago de Chile: Ed. Universidad de Concepción & Bravo y Allende.*

Ferrer-i-Cancho, R. y B. Elvevåg (2010). «Random texts do not exhibit the real Zipf's law-like rank distribution». En: *PLoS One* 5.3, e9411.

Foster, D. P., M. Liberman y R. A. Stine (2013). «Featurizing text: Converting text into predictors for regression analysis». En: *The Wharton School of the University of Pennsylvania, Philadelphia, PA.*

Jurafsky, D. y J. H. Martin (s.f.). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.*

Kant, I. (1978). «Crítica a la razón pura. Prólogo a la segunda edición». En: *Madrid. Alfaguara.*

Kelleher, J. D. y B. Tierney (2018). *Data science.* MIT Press.



## Bibliografía II

Montemurro, M. A. (2001). «Beyond the Zipf–Mandelbrot law in quantitative linguistics». En: *Physica A: Statistical Mechanics and its Applications* 300.3-4, págs. 567-578.

Muñoz-Basols, J., N. Moreno, T. Inma y M. Lacorte (2016). *Introducción a la lingüística hispánica actual: teoría y práctica*. Routledge.

Pinker, S. (2003). *The language instinct: How the mind creates language*. Penguin uK.

Suzuki, R., J. R. Buck y P. L. Tyack (2005). «The use of Zipf's law in animal communication analysis». En: *Animal Behaviour* 69.1, F9-F17.

Wikipedia (2023). *Ley de Zipf — Wikipedia, The Free Encyclopedia*.

<http://es.wikipedia.org/w/index.php?title=Ley%20de%20Zipf&oldid=148265153>. [Online; accessed 13-November-2023].

